



## Extraction de l'information implicite par analyse textuelle de sites Web en UNICODE

Bernard Dousset

### ► To cite this version:

Bernard Dousset. Extraction de l'information implicite par analyse textuelle de sites Web en UNICODE. Second séminaire VSST: Veille Stratégique Scientifique & Technologique, Systèmes d'information élaborée, Bibliométrie, Mar 2009, Nancy, France. hal-00379800

**HAL Id: hal-00379800**

**<https://hal.science/hal-00379800>**

Submitted on 5 May 2009

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# EXTRACTION DE L'INFORMATION IMPLICITE PAR ANALYSE TEXTUELLE DE SITES WEB EN UNICODE

**Bernard DOUSSET**, [dousset@irit.fr](mailto:dousset@irit.fr)

Institut de Recherche en Informatique de Toulouse

IRIT-SIG-EVI, Université Paul Sabatier, 118, route de Narbonne, 31062 Toulouse cedex 9 (France)

## **Mots-clés :**

UNICODE, langues étrangères, information implicite, signaux faibles, information semi structurée, fouille de données, fouille de texte.

## **Keywords :**

UNICODE, foreign languages, implicit information, weak signals, semi-structured information, data mining, text mining.

## **Palabras clave :**

UNICODE, lenguas extranjeras, información implícita, señales débiles, semi-estructuradas información, minería de datos, minería de texto.

## **Résumé**

Afin de diversifier les sources d'information que permet de traiter notre plate-forme "Tétralogie" dédiée à la veille stratégique, nous nous proposons de l'adapter, dans un premier temps, au traitement de la langue chinoise en nous basant sur son codage informatique en UNICODE. Il est alors possible de travailler sur des segments de texte repérés dans les différents dictionnaires disponibles (généraux ou spécifiques d'un domaine) et que nous pouvons alimenter par le traitement de bases d'information semi structurées et balisées en champs sémantiques : auteurs, mots-clés, organismes, journaux, ... Par la suite, des langues comme le japonais, le coréen, l'arabe seront abordées suivant le même principe, une exploitation multilingue pouvant même être envisagée via l'anglais ou un autre langage pivot.

## **Abstract**

In order to diversify the sources of information that can handle our platform "Tétralogie" dedicated to the business intelligence, we will adapt, as a first step, treatment of the Chinese language by using computer coding UNICODE. It is then possible to work on text segments identified in the various available dictionaries (general or specific domain) and we can feed the processor databases and semi-structured semantic marked fields: authors, words key agencies, newspapers, ... As a result, languages such as Japanese, Korean, Arabic will be addressed on the same principle, a multilingual exploitation can be seen via the English language or another pivot.

# 1. Problématique

Nous proposons une adaptation de notre plate-forme d'analyse stratégique ''Tétralogie'' au traitement des langues non européennes comme l'arabe ou le chinois dont le codage peut être ramené au système UNICODE UTF-8 en utilisant les capacités de transcodage d'outils comme *MS-Word*.

En se basant sur le codage informatique unifié de ces langues, nous pouvons détecter les entités nommées via le traitement de bases semi structurées comme les publications scientifiques, les brevets et la presse. Il est alors possible d'établir des dictionnaires assez complets de termes scientifiques et techniques (mots-clés, classifications, ...), de noms d'acteurs (auteurs, personnalités, entreprises, organismes publics, laboratoires, ...), de lieux géographiques, de dates et de les traduire dans la langue de l'utilisateur. Le croisement de ces briques signifiantes de texte doit nous conduire à l'établissement des réseaux sociaux et des réseaux sémantiques qui expliquent en grande partie les stratégies sous-jacentes du domaine. Si, de plus, le facteur temps est pris en compte (évolution des réseaux), nous pouvons avoir une vision prospective (tendances, trajectoires des acteurs, type d'évolution, ruptures technologiques, ...). Nous pouvons ensuite analyser le texte libre (titre, résumé, texte intégral) afin d'en extraire les nouveaux termes et donc les nouvelles tendances. Un premier traitement des textes à analyser permet de baliser l'ensemble des termes déjà identifiés (conscient collectif) et de les marquer comme entités indissociables (mots composés ou segments). Un second traitement, de type statistique, permet de découvrir de nouvelles séquences émergentes non présentes dans les dictionnaires initiaux mais représentant une nouvelle terminologie qui, si elle s'organise, permet de détecter les fameux signaux faibles (concepts émergents dans le domaine étudié). Nous appelons cela l'inconscient collectif, car il ne s'agit ni de mots-clés, ni de termes proposés par les auteurs eux-mêmes (free-terms), mais d'un vocabulaire nouveau souvent très précis et qui n'est utilisé que par une infime partie de la communauté (les personnes qui ont lu ou entendu des informations dignes d'intérêt et qu'elles reprennent à leur compte). Il se peut même que ces personnes ne se soient jamais concertées, et nous savons pourtant qu'elles valident, par leurs écrits, une idée émise dans un document (scientifique, technique ou d'actualité) ou lors d'une conférence. Il s'agit d'un consensus tacite, qui pourra être validé par croisement avec les signaux forts (connexions aux autres domaines) ou par identification de l'origine (qualité des acteurs concernés). Pour illustrer notre propos, nous allons présenter une analyse d'un corpus d'information extrait de la base scientifique chinoise CQVIP. Nous donnerons aussi quelques indications pour traiter de la même manière d'autres sources comme la base chinoise CNKI, le site d'information en arabe Al Jazeera ou une base coréenne en ligne : e-koreanstudies.com.

## 2. Préparation de la base scientifique chinoise cqvip.com

### 2.1. Principe général de la méthode

Cette base, accessible par l'URL <http://www.cqvip.com>, consigne une grande partie de la production scientifique chinoise. Citons pour exemple les 3 000 notices bibliographiques récupérées par mes étudiants de M1 de l'IUP SID, sur le domaine de l'aquaculture, pour la période 2004-2007. En fait, sans limite de date, plus de 8500 notices étaient disponibles sur cette base. Le codage des informations en UNICODE, nous permet de travailler sur des séquences d'idéogrammes qui correspondent soit à des mots-clés, soit à des acteurs du domaine (journaux, colloques, organismes, laboratoires, auteurs, ...). Les passages en texte libre peuvent à leur tour être traités (Titre et Résumé) afin d'y détecter de nouvelles séquences souvent inconnues des experts du domaine. Leur traitement par agrégation dans des réseaux sémantiques permet d'en déduire la consistance (nombre de documents y faisant référence), l'actualité (apparition récente), la pertinence (domaine connexe et qualité des acteurs), la structure sémantique (nombre de segments émergents et typologie de leurs relations : verbes). On peut ainsi, sans parler la langue, extraire tous les éléments statistiques utiles à la prise de décision et même détecter les signaux faibles, qui représentent l'essentiel de la demande en terme d'information à haute valeur stratégique.

Dès la page d'accueil du site cqvip.com, un moteur de recherche est disponible, il doit être alimenté par des requêtes en chinois qui peuvent être traduites depuis l'anglais par les outils linguistiques de *Google*. Il suffit ensuite de cliquer sur la loupe pour obtenir les premières notices (certains champs sont masqués). Plusieurs options sont alors possibles : soit récupérer les notices visualisées par un copier-coller vers *MS-Word*, soit demander un téléchargement de l'ensemble des champs, soit encore demander à un robot de faire le travail.

Voici le protocole de travail que nous préconisons pour réaliser des analyses stratégiques depuis cette base documentaire.

- établir une équation de recherche en anglais et la traduire en chinois [*Google*]. Par exemple : ‘aquaculture’ devient ‘水产养殖’
- soumettre cette équation au moteur de recherche de la base cqvip.com,
- pour chaque page de résultat, la copier-coller dans *MS-Word 2003* ou dans *Open Office* <sup>1</sup>,
- enregistrer le fichier obtenu au format html,
- l'ouvrir avec *Internet Explorer*,
- afficher le code source de la page,
- l'UNICODE apparaît en mode texte ASCII sous la forme « &#12345; »,
- réenregistrer le code source en format texte.

Ce dernier fichier, après un reformatage par un programme *Perl* (détection des balises et des idéogrammes codés UNICODE voir figure 2), peut être entièrement analysé par notre plate-forme « *Tétralogie* » :

- création des dictionnaires (auteurs, mots-clés, journaux, organismes, villes, dates et chaînes répétées d'idéogrammes issues du texte libre),
- croisement des variables (tableaux 2D ou 3D binaires, de contingences ou de cooccurrences simples ou multiples),
- visualisation des résultats (tris de matrices, cartes factorielles, arbres de classification, dessins de graphes, cartes géostratégiques),
- détection des signaux faibles, ...

Une fois l'analyse stratégique effectuée, il est possible d'exporter les croisements dans une base de données *MySql* qui peut être interrogée via le Web [Ghalamallah 2008]. Si tous les segments de texte (nom des auteurs, mots-clés, journaux, ...) sont traduits du chinois à l'anglais (par *Google* ou *Systran*), l'utilisateur distant travaillera de façon transparente dans cette langue. Il lui sera alors possible de se focaliser, par filtrage multi critères, sur un point de détail qui l'intéresse particulièrement et d'obtenir toutes les statistiques possibles sur cet extrait des données sources (tops n, histogrammes, secteurs, réseaux d'acteurs, réseaux sémantique, évolutions, cartes géographiques, classifications, ...).

---

<sup>1</sup> On se sert, ici, des capacités de transcodage de ces outils pour passer de polices chinoises diverses à l'UNICODE. *MS-WORD 2007* n'est pas recommandé car il sait travailler avec plus de polices, il ne fait donc plus le transcodage voulu : « qui peut le plus peut le moins ».

用户名:  密码:  [登录](#) [IP登录](#) [用户注册/忘记密码](#) [电子期刊阅览室](#) [充值中心](#) [Google 学术](#) [旧版入口](#)

**VIP 维普资讯**

题名/关键词  搜尽8000余种期刊文章..... [文章搜索](#)

• 高级检索 • 常见问题

**中文期刊·专业文章** [首页](#) [期刊导航](#) [知识社区](#) [学者空间](#) [学术机构](#) [专题导航](#) [充值中心](#)

[临床医学](#) [财经](#) [中医中药](#) [教育](#) [化学工程](#) [农学](#) [自动化计算机](#) [生物学](#) [药学](#) [材料科学](#) [论文翻译](#) [论文发表](#)

**2007非常盘点：民生关注热门榜**

**学科分类**

**医药卫生**  
 临床医学 基础医学  
 预防医学 中医中药  
 药理学

**工程技术**  
 矿山工程 石油工业  
 冶金工程 材料科学  
 机械仪表 能源动力  
 电气工程 电子通信  
 化学工程 建筑科学  
 水利工程 交通运输  
 航空航天 环境安全  
 自动化计算机  
 纺织 食品

**自然科学**  
 生物 化学 力学  
 数学 物理 天文地球

**农林牧渔**  
 畜牧兽医 水产渔业  
 农学 林学

**人文社科**  
 管理 财经 政治  
 法律 哲学 宗教  
 军事 教育 文化  
 体育 艺术 语言  
 文学 历史 地理

**订阅服务**  
 电子期刊订阅  
 学科分类订读  
 纸版期刊订购  
 在线充值中心  
 阅读充值资费

**论文发表/论文写作**  
 专业提供论文发表和写作服务  
 服务保发过 010-63397249  
 QQ: 278123888

**LPG——后石油时代“枯木逢春”**

作为车用清洁替代燃料中技术最成熟、最容易推广使用的LPG(液化石油气)，近十年的推广运用却经历了一个曲折的发展过程，没能如人愿地顺利发展。在我们大力建设节约型社会的今天，关注和推广成熟替代燃料无疑是调整能源结构，改善大气环境，实现可持续发展的有效措施。

• 警惕不典型的干燥综合征 干燥综合征  
 • 工程轮胎生产现状与发展前景 工程轮胎  
 • 垂直管理模式下权力的配置与制约 垂直管理  
 • 中央与地方关系法治化基本问题研讨 中央与地方  
 • 我国私募证券投资基金“收益保底”探析  
 • 个人投资理财渠道大盘点 理财渠道  
 • 对有效市场假说的再认识 资本市场  
 • 伟大的资本博弈 资本博弈  
 • 社会舆论监督的法律保障问题研究 舆论监督  
 • “戴总”没有希望——国际舆论纷纷质疑赖赖嘛

**订阅电子期刊 建立个性化阅览室**

[下载](#) [PDF阅读器](#) [阅读全文需先安装](#) [简体中文版PDF阅读器](#)

**学者推荐** [更多学者>>](#)

**张靖周**  
 所在的机构 南京航空航天大学  
 主要研究方向 1. 传热传质学  
 2. 燃烧学 3. 红外隐身

• 直升机红外抑制器两种进气方式下的实验研究  
 • 压电驱动狭缝微孔自耦合射流流动特征研究  
 • 爆震管内波与火焰相互作用机理的试验  
 • 压电驱动自耦合射流流动特性实验  
 • 突片射流冲击冷却换热特性的实验

**【水利工程】** 倪晋仁 杨志峰 黄润秋 陈守煜  
**【化 学】** 俞善信 高廉 寇元 张克武  
**【财 经】** 王靖慈 张维迎 杨开忠 魏守华  
**【教育研究】** 张尧学 谈松华 王冀生 何克抗  
**【天文地球】** 王绍武 郑荣才 付广 施雅凤

**学术机构** [更多>>](#)

**【浙江】** 浙江大学附属第二医院  
**【云南】** 昆明冶金研究院  
**【广西】** 广西师范大学  
**【湖北】** 中国长江三峡工程开发总公司  
**【江西】** 东华理工大学

**快乐学习，下一站？ 灌水专区 -->**

**知识社区**

水产养殖

☐ 全选记录 ☐ 下载记录 ☐ 打印记录 ☐ Email记录 按时间筛选: [全部](#) 显示方式: [文摘显示](#)

1. [标题] **日本对虾精养高产技术初探**  
 [作者] 周晨光 陈佳颖 黄秀荣  
 [机构] 浙江省舟山市定海区海洋与渔业局, 316000  
 [文摘] 舟山市绿源水产养殖公司2005年采用围塘塑料大棚进行日本对虾精养高产技术研究, 每667m<sup>2</sup>养殖产量达到366.87kg, 经济效益十分显著。本文主要介绍该公司围塘日本对虾精养高产技术, 供参考。  
 [刊名] >>>齐鲁渔业-2008;25(2)-27-27

[相关文献](#)

2. [标题] **微生态制剂及其在水产健康养殖中的应用**  
 [作者] 潘小红 [1] 陈国贵 [1] 徐学峰 [2]  
 [机构] [1]河南省平顶山市农业技术推广站, 467000 [2]平顶山市水产技术推广站  
 [文摘] 近年来, 在水产养殖业中, 微生态制剂作为绿色饲料添加剂、水质改良剂以及对鱼类健康、预防疾病、促进生长和品质改善所起的显著作用, 越来越被人们所重视。并以其无毒副作用, 无耐药性, 无残留污染, 效果显著等特点逐渐得到广大水产养殖者的认可, 不少地方把目光注现在微生物技术在水产养殖的应用上来, 利用微生物制剂的辅助作用建立水产健康养殖模式, 实现无公害化养殖。  
 [刊名] >>>齐鲁渔业-2008;25(2)-50-52

[相关文献](#)

3. [标题] **高青县渔业生产驶入“高速路”**  
 [作者] 李恒水 史春来  
 [文摘] 高青县渔业生产进入快速发展阶段。2007年全县水产养殖面积达到3400hm<sup>2</sup> (5.1万亩), 水产品总产量2.6万t, 产值达2.2亿元。发挥了水产业在全县农业中的四大支柱的作用, 被省海洋与渔业厅列入全省水产养殖重点县。  
 [刊名] >>>齐鲁渔业-2008;25(2)-61-61

[相关文献](#)

4. [标题] **禹城渔业污染源清查摸底全面完成**  
 [作者] 尹国存  
 [文摘] 最近, 禹城市水产局根据《第一次全国污染源普查清查工作细则》、《山东省水产养殖业污染源普查实施方案》的要求, 认真落实省海洋与渔业厅《关于开展水产养殖业污染源清查工作的通知》精神, 成立专门班子, 强化工作措施, 制定清查方案, 确保人力、物力、车辆到位, 向各乡镇下发通知, 召开工作会议部署, 按要求深入到户,  
 [刊名] >>>齐鲁渔业-2008;25(2)-63-63

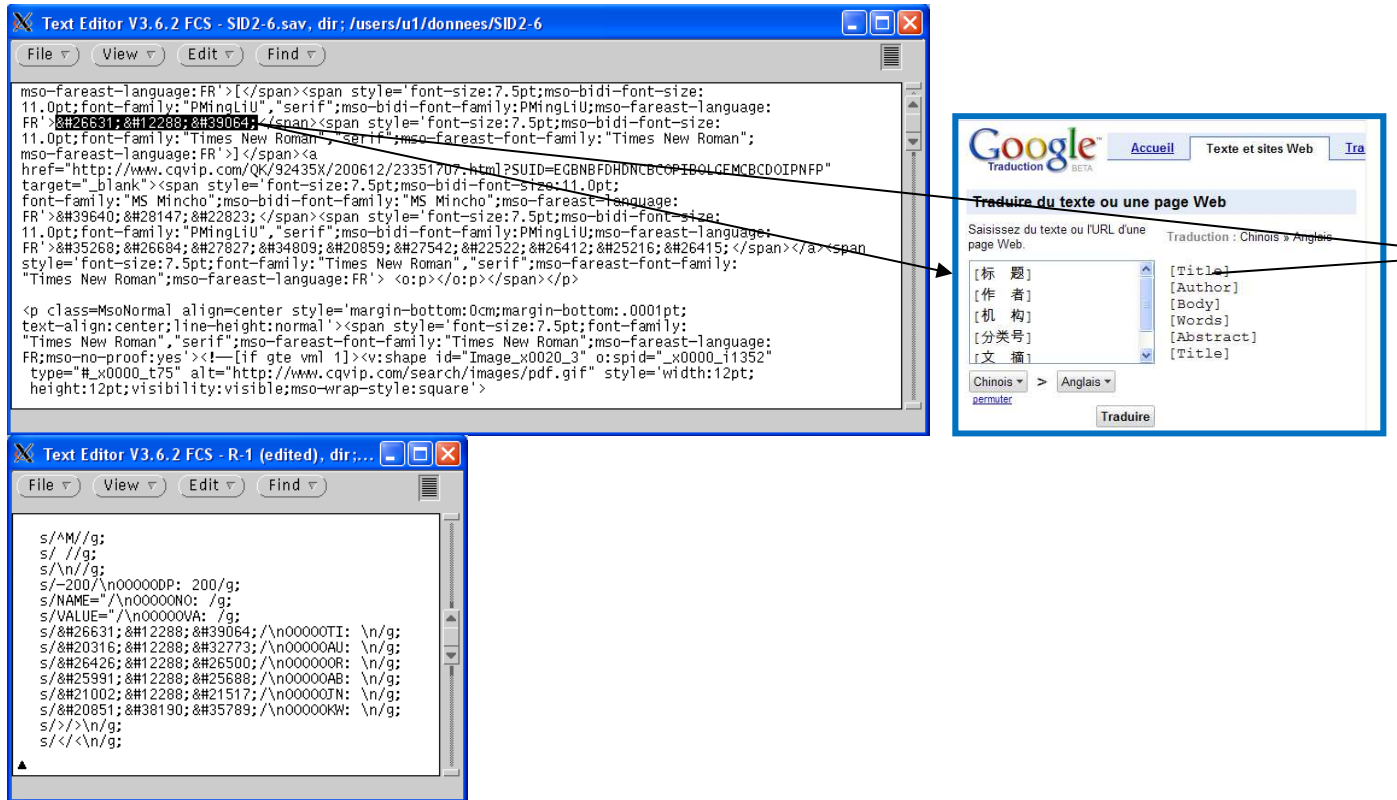
[相关文献](#)

5. [标题] **无棣县狠抓水产品品质**  
 [作者] 邵红梅 温孟泉  
 [文摘] 山东省水产品质检检测中心对无棣县水产养殖示范区对虾、鳊鱼、梭子蟹等近10个水产品种抽样监测化验表明, 各项化验指标均高于国家标准, 产品合格率100%。  
 [刊名] >>>齐鲁渔业-2008;25(2)-64-64

[相关文献](#)

Figure 1 : le site cqvip.com (en haut le moteur de recherche, à droite un extrait du résultat)

## 2.2. Reformatage de la base



*Figure 2 : Reformatage de la base CQVIP : traduction par Google et transformation des balises par un utilitaire Perl*

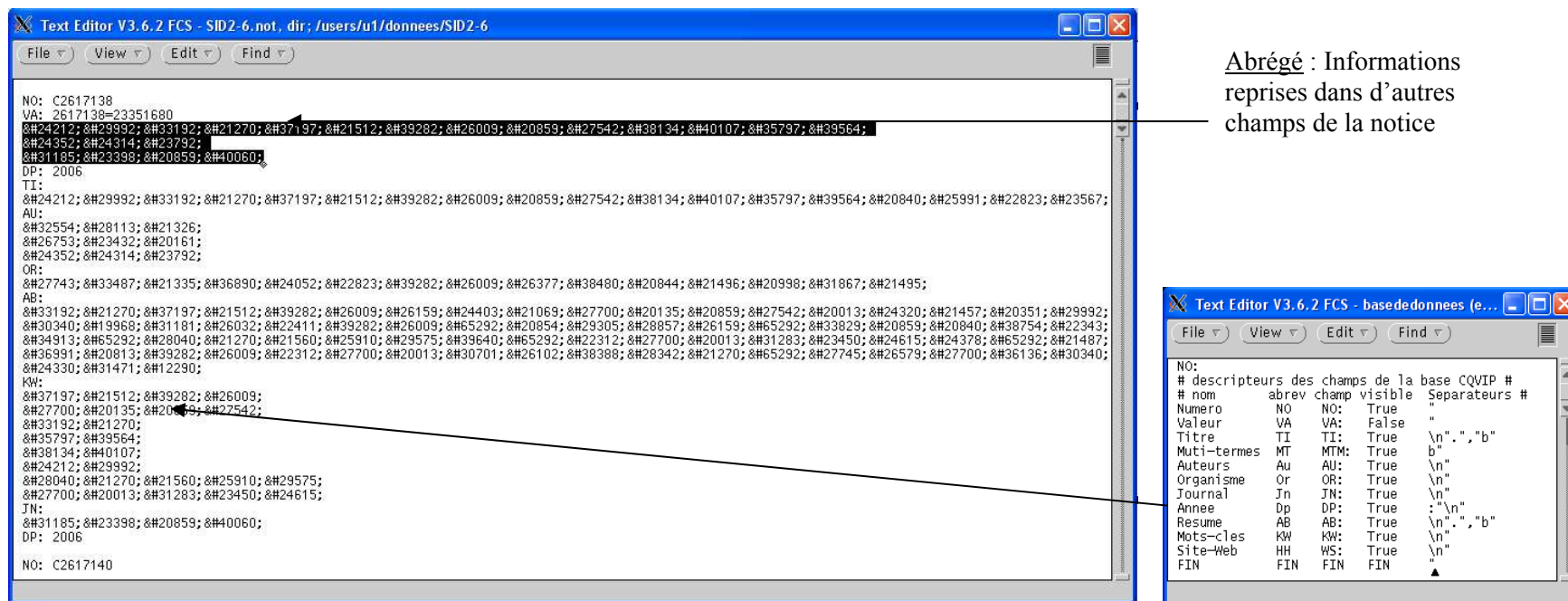
Le but de cette étape est multiple :

- éliminer la mise en forme du texte (html) qui n'apporte rien au contenu, mais qui représente 90% du poids du fichier,
- reconstituer des chaînes de texte qui sont séparées en tronçons par la mise en forme. Cette étape est nécessaire car de nombreux mots clés, par exemple, sont coupés (changement de police au niveau html). Est-ce volontaire pour générer l'analyse automatique et l'indexation ?
- rebaliser le texte par des balises ASCII (du type : TI:, AU:, OR:, ...) traduites des balises existantes en chinois,
- certaines balises et leur contenu ne sont pas visibles sur la page web, mais sont présentes dans le code source en ASCII (ex : KW:)
- rajouter des balises inexistantes dans le texte en les créant à partir de la mise en forme html : DP:, NO:, VA:,



- conserver les informations qui sont codées en caractères latins ou en chiffres arabes comme les dates, les numéros ou certains termes comme les auteurs occidentaux, les formules ou les éléments techniques,
- éviter les doublons dus à la présence simultanée de deux versions du même terme (ex : noms des auteurs cliquables sur la page web).

Une fois le reformatage/rebalisage terminé, le corpus se trouve sous une forme analysable par « *Tétralogie* » au même titre qu'une base initialement codée en ASCII comme *PASCAL*, *FRANCIS*, *INSPEC* ou *SCI*. Des méta données compatibles avec ce balisage vont permettre ensuite de piloter les automates qui vont traiter le texte : extraction des items de tous les champs balisés, puis croisement des champs entre eux afin d'établir des tableaux de cooccurrences à 2 ou 3 dimensions.



**Figure 3 :** Une notice bibliographique reformatée (balises en ASCII, texte chinois en UNICODE) et ses méta données

Dans la figure 3, nous pouvons remarquer les balises de chaque champ en ASCII et le texte (contenu) en UNICODE. Dans la notice C2617138, le titre, le premier auteur, le journal et la date forment un abrégé en début de document, ces informations sont ensuite reprises dans les champs correspondants : TI:, AU:, JN:, DP:. Le champ VA: ne sera donc pas utilisé dans l'analyse.

Si nous analysons visuellement le document, il comporte 3 auteurs (de 3 idéogrammes chinois = 3 codes), un seul organisme, 8 mots clés (ici de 2 à 5 idéogrammes) plus le journal et la date : 2006. Nous allons voir par la suite que le titre et l'abstract feront l'objet d'un traitement sémantique particulier : détection des chaînes répétées d'idéogrammes ne correspondant à aucun mot clé, d'où possibilité d'une terminologie qui n'apparaît pas dans l'indexation proposée.

Les méta données (à gauche) décrivent le format obtenu : nom complet de chaque champ, abréviation de ce même champ (sert à nommer les fichiers et répertoires associés à son analyse), identifiant exact du champ dans la notice (ex : TI: pour le champ Titre), TRUE pour l'utiliser dans l'analyse, suite des séparateurs utilisés pour découper le texte (chaîne de caractère, "b " pour l'espace, "\n" pour la fin de ligne, "ORD2" pour le 2° segment, ...). Dès que nous disposons du corpus reformaté et de ses méta données, il est possible de l'analyser par les mêmes techniques que pour une source en alphabet latin, y compris réaliser son traitement sémantique et proposer une indexation plus fine et récente.

## 2.3. Problèmes de traduction

```

&#26446;&#23159; Li Ting
&#26446;&#21516;&#24198; Li Tongqing
&#27754;&#33673; Li Wang
&#26446;&#21169;&#24180; Li was
&#26446;&#21033;&#21355; Li Wei-1
&#26446;&#24013; Li Wei-2
&#26446;&#38686; Li Xia-1
&#40654;&#22799; Li Xia-2
&#26446;&#21521;&#32676; Li Xiangqun
&#26446;&#20808;&#26126; Li Xianming
&#26446;&#27427; Li Xin
&#26446;&#24314;&#26143; Li Xing-1
&#26446;&#26143; Li Xing-2
&#26446;&#23398;&#33391; Li Xueliang
&#26446;&#23398;&#20184; Li Xue-pay
&#26446;&#23398;&#36805; Li Xun
&#26446;&#24433; Li Ying
&#26446;&#21191; Li Yong
&#26446;&#21191;&#24070; Li Yong-fan
&#26446;&#27704;&#27665; Li Yongmin
&#26446;&#29577; Li Yu
&#26446;&#21407; Li Yuan
&#26446;&#27835;&#20853; Li Zhi-bing
&#26446;&#21338;&#36995; Li Zongsui
&#36830;&#24120;&#24179; Lian Chang Ping
&#26753;&#20840;&#25991; Liang full text
&#26753;&#21073;&#24179; LIANG Jian-ping
&#26753;&#22025;&#19968; LIANG Jia-one
&#26753;&#33804;&#38738; LIANG Meng-qing
&#26753;&#22825;&#32418; Liang Tian-hong
&#26753;&#23398;&#27665; Liang Xue-min
&#20237;&#22025;&#20142; Liang-1
&#23731;&#32768;&#20142; Liang-2
&#32735;&#26093;&#20142; Liang-3
&#24278;&#28023; Liao
&#24278;&#26124;&#23481; LIAO Chang-rong
&#24278;&#26757;&#26480; LIAO Mei-jie
&#24278;&#21151; Liao Zhigong
&#35299;&#25918; Liberation
&#29579;&#28872;&#21326; Lie-hua
&#40654;&#20809;&#31243; Liguang Cheng
&#33539;&#31435;&#27665; Li-Min Fan
&#40654;&#26126; Liming
&#26519;&#26149;&#21451; Lin Chun-you
&#26519;&#21551;&#23384; Lin Cun
  
```

```

24 CHEN CHANGFU
18 CHEN AI-PING
13 MENG CHANG-MING
12 WANG WENBIN
10 XIONG YAN CHENG
9 PAN JIN-MING
9 &#28504;&#31168;&#33714;
8 SONG CHANG-TAI
8 LI-3
8 CHANG KIM
7 XIAO PEI PHILIP
7 SUN KE-YEAR
7 LIU JIN
7 DE-AN ZHAO
6 ZHOU NEW
6 ZHANG WEN-GE
6 XIE GANG-1
6 XIANGYANG
6 SHANGHAI RED
6 LI TONGQING
5 ZOU YONG
5 ZHU XUEBAO
5 ZHOU KAI
5 YAO WEIZHI
5 XU ZIRONG
5 WANG KAI-YU
5 TAN HONG NEW
5 SHAO QING ARE
5 LUO ZHI
5 LIU WENBIN
5 &#29579;&#24422;&#27874;
5 &#26446;&#27491;&#39134;
4 ZHU WEN
4 ZHANG ZHONGYUAN
4 YI
4 YANG LING
4 XIAO YUAN-JIN
4 XI LIU WEI
4 WUJIANG
4 WU-SHENG
4 WANG GUANG-JUN
4 SHU-LIN WANG
4 PAN LIN TAK
4 MIAO XIANGWEN
4 LIU XUN
  
```

Afin de décoder l'UNICODE (et donc, ici, le chinois), nous avons établi des dictionnaires de correspondances entre les noms d'auteurs en chinois et leur traduction en phonétique (Pinyin) grace au traducteur de *Google*. Mais dans ce cas, se posent deux problèmes :

- *Google* n'arrive pas à traduire tous les noms et restitue alors l'UNICODE (voir 7° auteur)
- Plusieurs auteurs de codes différents peuvent avoir la même correspondance, d'où une ambiguïté très néfaste à l'établissement de réseaux sociaux pertinents et la nécessité impérieuse de corriger ce problème.

Nous avons choisi, dans le premier cas, de garder les codes et, dans le second, d'ajouter un différenciateur numérique aux noms traduits en totalité ou partiellement (ici, LI-3 et XUE GANG-1, par exemple).

La lecture des dictionnaires, des cartes, des graphes, des dendogrammes est ainsi plus parlante, tout en évitant d'introduire des artéfacts dus à la traduction approximative souvent proposée par *Google*. Dans ces conditions, il est possible de faire la traduction avant l'analyse, sinon il faut tout analyser en chinois (donc en UNICODE, c'est-à-dire en aveugle) et ne traduire qu'à la fin, afin de restituer un résultats fiable et intelligible (traduit) pour les utilisateurs. Deux noms identiques peuvent alors cohabiter dans la même visualisation du moment qu'ils ont été différenciés par leurs codes respectifs tout au long de l'analyse. Les deux Li-Xing (-1 et -2) seront alors différenciés non par leur nom (homonymie complète après traduction) mais par leur position dans les cartes ou leur appartenance à des classes différentes. On retrouve le problème bien connu de

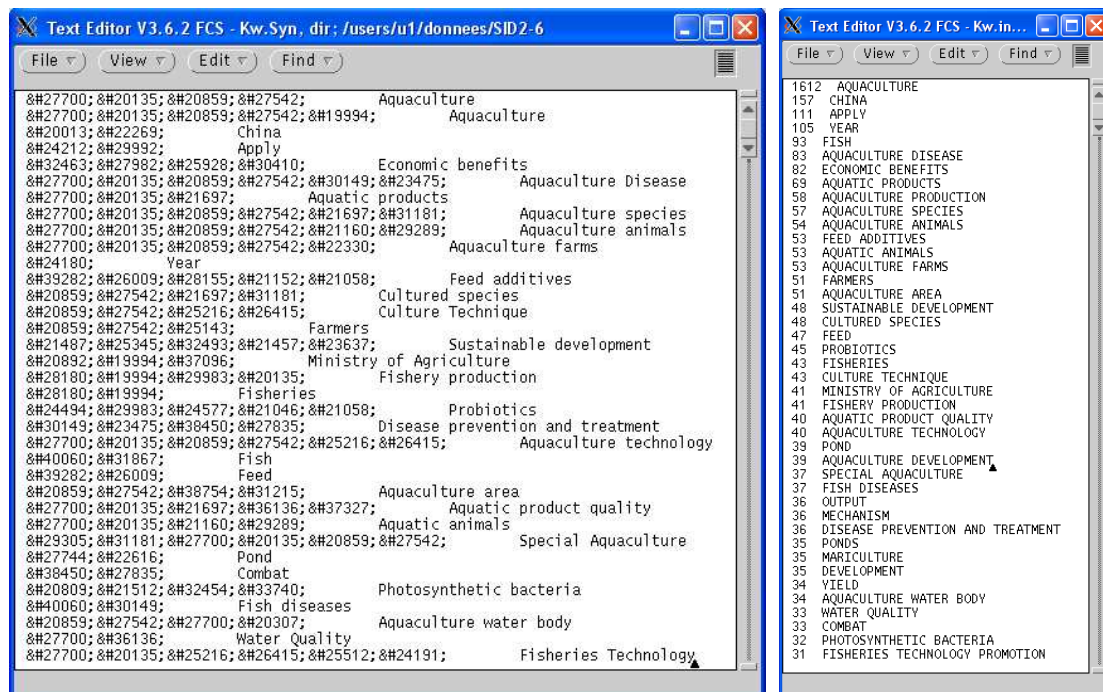


l'ambiguïté des noms asiatiques, mais ,ici, il est résolu presque'entièrement à la différence des bases occidentales où il perdure.

**Figure 4 : Correspondance UNICODE-Pinyin et occurrences**

Un autre problème de traduction peut se poser pour la terminologie technique (mots clés, indexation additionnelle, texte intégral) car *Google* ou *Systran* ont des difficultés sur ce point là. Absence dans leurs dictionnaires (termes trop techniques ou trop récents), problème de contexte, complexité des phrases, ambiguïté du discours. La plupart du temps, cette incertitude est levée, car le nombre de termes bien traduits, dans par exemple un cluster sémantique, est suffisant pour comprendre la nature du concept mis en valeur et sa pertinence. Bien entendu le recours à un expert connaissant bien la langue est recommandé.

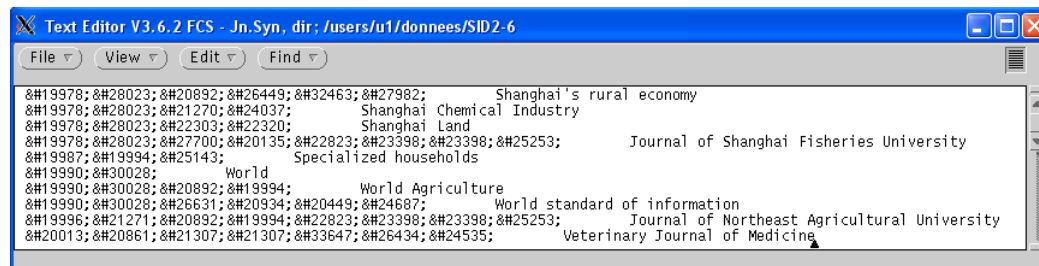
- Pour les mots-clés, le problème est similaire. En effet, certains mots-clés, qui en UNICODE sont différents, se voient attribuer la même traduction par *Google*. Ce phénomène est heureusement assez rare, ce qui ne compromet pas trop l'interprétation des experts.



**Figure 5 : traduction des mots clés et leurs occurrences calculées après traduction.**

- Pour les journaux, la traduction est sans souci bien que parfois un peu folklorique.

Ci-contre, le début du dictionnaire de synonymes donnant la correspondance entre mots clés chinois en UNICODE et leur traduction *Google* en anglais. Les occurrences des mots clés sont ensuite calculées pour l'anglais, d'où fusion des termes traduits de la même façon. Ici, dans notre exemple, le terme générique le plus fréquent « aquaculture » cumule les occurrences de plusieurs formes, c'est beaucoup moins gênant que dans le cas des auteurs, mais on risque de perdre certaines nuances ou faire des approximations peu judicieuses. A souligner que dans l'affichage html des documents dans *Internet Explorer* (figure 1), le champ mot-clé n'apparaît pas, il est simplement masqué par la mise en forme, mais bien présent dans le code source et donc restitué à l'issue du reformatage (figure 3). Ces mots clés seront très utiles dans la phase de traitement sémantique du titre et du résumé, car ils imposeront des chaînes d'idéogrammes ayant un sens dans le contexte étudié.



- Par contre, pour les organismes, plusieurs formes peuvent être rencontrées (problème inhérent à la morphologie des adresses). D'où la nécessité de mettre au point un dictionnaire de synonymes permettant de corriger et donc de regrouper toutes les formes équivalentes rencontrées.

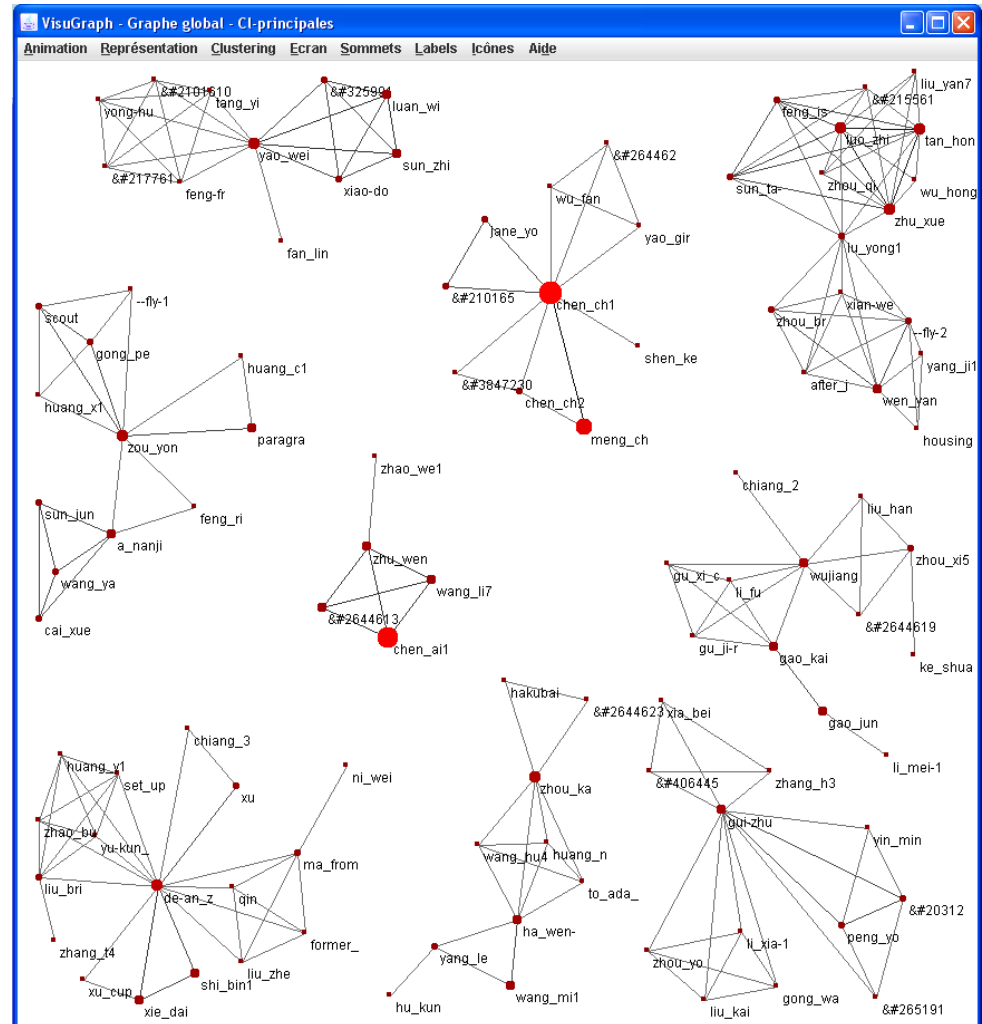
*Figure 6 : traduction des journaux.*

## 3. Analyse de l'aquaculture en Chine

### 3.1. Les réseaux d'acteurs

Après avoir traduit les noms des auteurs en anglais et avoir différencié ceux dont les noms sont différents en chinois et identiques en anglais, nous avons croisé les auteurs ayant au moins deux publications (connecteurs potentiels). La figure 7 présente la topologie des principales équipes. Première constatation : très peu de collaborations affichées dans les publications scientifiques chinoises. Seconde constatation, les équipes sont le plus souvent dirigées par un mandarin qui contrôle 2, 3 ou 4 sous équipes distinctes. Certains noms ne sont pas traduits, d'autres ont des traductions folkloriques, car leur nom a une autre signification en chinois :

- &#21476;&#32676;&#32418; Ancient group of red
- &#37329;&#24425;&#26447; Apricot Jincai
- &#21556;&#26089;&#20445; As early as Paul Wu
- &#23391;&#21644;&#24179; Bangladesh peace
- &#34013;&#27491;&#21319; Blue is up
- &#21830;&#24503;&#31456; Business ethics chapter
- &#21830;&#19975;&#25104; Business Wancheng
- &#34081;&#31168;&#20029; Cai beautiful
- &#34081;&#24314;&#22564; Cai embankment
- &#38472;&#22269;&#20820; Chan Kwok-rabbit
- &#31456;&#31179;&#34382; Chapter autumn tiger
- &#38472;&#26435;&#20891; Chen the right to military
- &#37011;&#27491;&#33829; Deng Zhenglai business
- &#30224;&#33673;&#33805; Die in a prison Liping
- &#21035;&#25991;&#32676; Do not text-qun

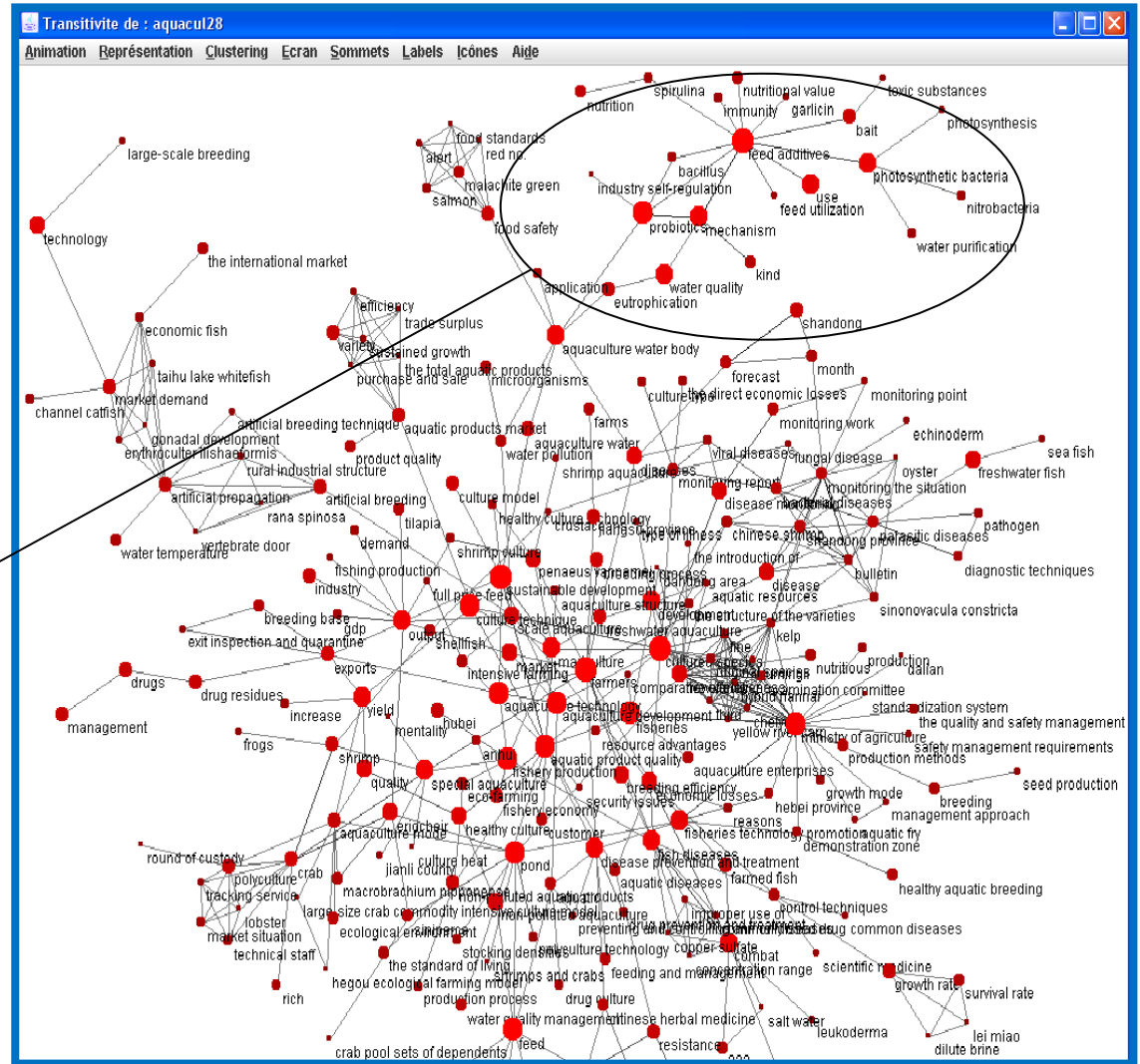


posés dans les notices et ainsi faire apparaître

Voici l'exemple d'un des clusters trouvés :

- Nutrition
- Spirulina
- Nutritional value
- Immunity
- Garlicin
- Bait
- Toxic substances
- Photosynthetic bacteria
- Photosynthesis
- Nitrobacteria
- Water purification
- Feed utilization
- Bacillus
- Probiotic
- Industry self-regulation
- Mechanism

128



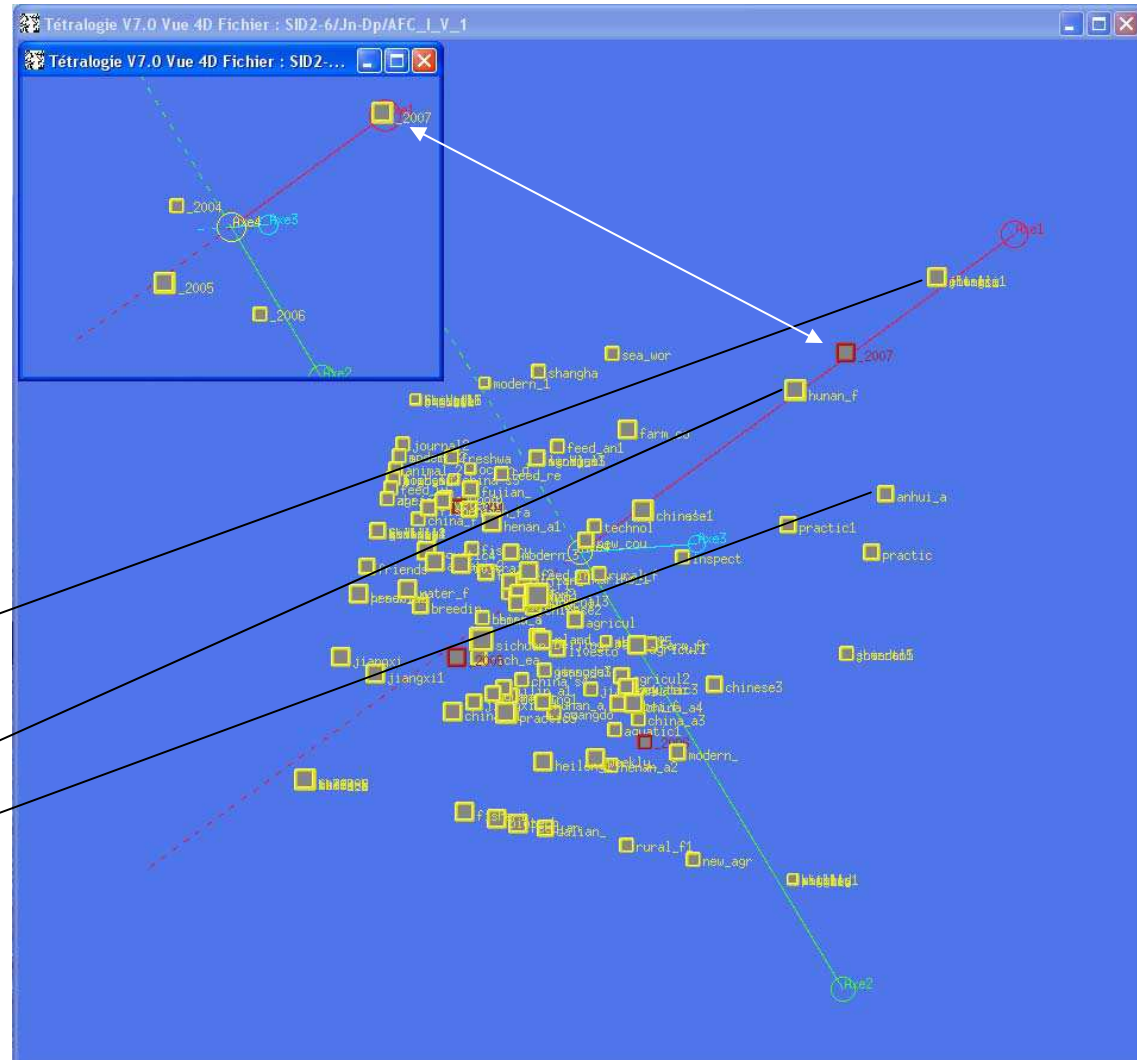
- Kind
- Water quality ...

*Figure 8 : réseau sémantique des mots clés proposés dans CQVIP.*

### 3.3. Analyse de l'évolution

Pour illustrer une première méthode, nous avons choisi d'analyser le profil d'évolution des journaux sur les quatre années de notre étude, à savoir [2004 ; 2007]. Une analyse factorielle des correspondances sur la matrice de cooccurrences Journaux X Dates (Jn X Dp) nous permet de visualiser facilement (figure 9) les différents profils sur un tétraèdre régulier représenté, ici, en 3 D. Au niveau de chaque sommet les journaux qui ne sont présents qu'un an : les nouveaux journaux sont donc en haut à droite. Sur une arête ceux qui ne sont présents que sur 2 années : ceux de 2006 et 2007 sont sur l'arête de droite. Ceux présents 3 ans sont sur une face, ceux présents tout le temps sont à l'intérieur du tétraèdre et attirés par l'année où ils sont le plus présents. La carte en haut à gauche ne visualise que les années, l'autre y ajoute les journaux. Ces deux cartes peuvent être orientées dans tous les sens, l'azimut de visualisation choisi peut être exporté sur l'autre carte ou vers toute carte visualisée sur l'ordinateur d'un utilisateur distant. On peut ainsi lui communiquer un point de vue significatif comme, ici, celui servant à identifier les nouveaux journaux abordant l'aquaculture.

- Chemical friends
- China education information
- Jiangsu health service management
- China electrical industry
- Intelligent buildings and urban information
- Electric locomotive and mass transit vehicles
- China metallurgical
- Anhui agricultural science
- Hunan feed
- Shanxi agricultural: the village committee director

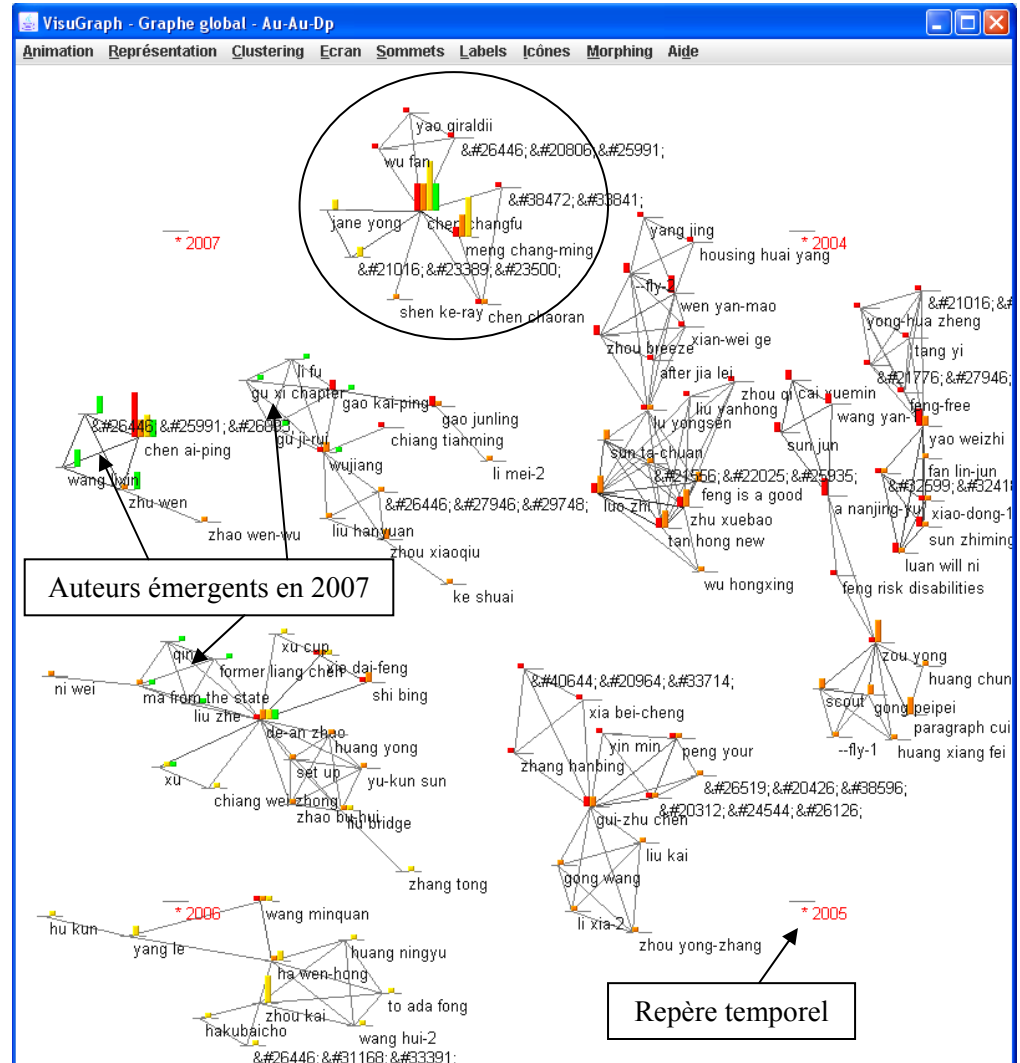




- China inspection and quarantine

Une seconde méthode consiste à réaliser un tableau de cooccurrence visualiser le réseau évolutif de ces auteurs sur un graphe adapté. L sur une horloge. Les auteurs sont attirés par ces sommets fictifs et viennent se placer vers le centre du graphe s'ils sont présents tout le temps, vers le repère correspondant s'ils ne sont présents qu'une fois, ou dans une position intermédiaire s'ils sont présents au cours de plusieurs périodes consécutives ou non. Cette analogie espace-temps fonctionne un peu comme la carte factorielle précédente, mais on peut mixer cette approche avec le dessin classique de graphe (sans repère additionnel). Nous obtenons alors, comme ci-contre, un graphe qui fait apparaître les principales équipes avec leurs évolutions respectives. L'histogramme coloré attaché à chaque sommet nous indique son évolution quantitative. La position vis-à-vis de ses collaborateurs indique l'époque de son implication dans l'équipe. Ses liens précisent avec qui et combien de fois il a collaboré. Dans la figure 10, nous avons rassemblé les dynamiques des principales équipes chinoises dans le domaine de l'aquaculture. Certaines sont persistantes, d'autres émergentes, d'autres ont cessé momentanément ou définitivement de publier dans le domaine. Il est facile de repérer les leaders (taille des icônes), de détecter les nouveaux auteurs (en vert), les auteurs qui ne sont plus présents (pas de vert), les auteurs clés, connecteurs entre sous équipes. Par exemple, dans l'équipe du haut, seul le leader (Chen Changfu) reste encore actif, il avait un gros collaborateur jusqu'en 2006 (Meng Chang-Ming), il a dirigé deux équipes distinctes en 2004, il a travaillé avec Shen Ke-Ray en 2005 et une équipe de 2 personnes en 2006. Par contre, les trois équipes de gauche ont de nombreux auteurs émergents et des leaders persistants. Les autres équipes ont disparu, les quatre de droite dès 2006. Enfin, il est possible de réaliser une analyse factorielle multiple (AFM) sur ce même tableau 3D (Au-Au-Dp) et faire apparaître les trajectoires des auteurs s'ils changent d'équipe. Ce n'est pas le cas ici, donc pas de mobilité constatée sur les quatre ans de notre étude.

**Figure 9 : analyse factorielle sur la matrice Journaux X Dates**





**Figure 10** : réseau montrant l'évolution des principales équipes.

### 3.4. Analyse sémantique du texte libre

Pour analyser le texte libre, nous utilisons le dictionnaire de mots-clés de la figure 5 (termes techniques, mots-clés), plus un dictionnaire de mots vides (les mots outils de la langue : of, the en anglais), et éventuellement un dictionnaire de synonymes (équivalences terme à terme ou terme avec concept générique : car, vehicle). Le texte libre est, dans un premier temps, découpé en phrases grâce aux ponctuations présentes dans le titre et surtout dans le résumé. Les chaînes d'idéogrammes correspondant aux mots-clés sont ensuite recherchées dans le texte et elles sont complétées par la détection de nouvelles chaînes donc le nombre d'occurrences dépasse un seuil défini par l'utilisateur en fonction de la taille du corpus. Ces nouvelles chaînes d'idéogrammes qui intègrent éventuellement des mots-clés existants sont traduites afin d'essayer d'en comprendre le sens. Il y a alors deux possibilités, ou *Google* en donne une traduction acceptable (mot-clé manquant dans l'indexation de la base, ici CQVIP), ou la traduction ne veut rien dire (nouveau concept absent du dictionnaire de *Google* et donc certainement émergent). Dans ce second cas, il faut revenir au contexte d'utilisation de cette nouvelle chaîne d'idéogrammes et recourir à un spécialiste du domaine, ici l'aquaculture, et parlant le chinois. Une autre possibilité, est de voir si ces nouveaux segments forment des clusters [ROUX 2009] (cooccurrences de plusieurs d'entre eux dans quelques documents). Il est alors fort possible que nous ayons détecté un signal faible (cohérence, simultanéité, consensus). Pour le valider, il suffit de le croiser avec tous les autres champs de la base (Auteurs, Organismes, Mots-clés, Journaux, Dates). S'il est récent, dans un bon journal, dans une bonne équipe et connexe à des domaines connus (mots-clés), il a certainement du potentiel (ce qui reste à valider auprès d'un expert en lui donnant à analyser tous les documents contribuant à ce signal). Nous pouvons ainsi, sans connaître la langue, détecter une information implicite présente dans le corpus et inaccessible par la simple lecture. Ce travail de détection des signaux faibles est en fait très demandé par les décideurs car il correspond à un besoin naturel de détecter précocément l'innovation afin d'évaluer la pertinence de nouvelles pistes et de réagir en conséquence. Ci-dessous, une liste de termes pertinents détectés (nouvelles chaînes d'idéogrammes) et un cluster sémantique émergent.

&#20859;&#27542;&#22616;	Breeding pond	繁殖塘
&#20859;&#27542;&#21487;&#25345;&#32493;&#21457;&#23637;	Sustainable development of aquaculture	养殖可持续发展
&#20859;&#27542;&#25345;&#32493;&#20581;&#24247;&#21457;&#23637;	Sustained and healthy development of aquaculture	养殖持续健康发展
&#20859;&#27542;&#27827;&#34809;	Breeding crab	养殖河蟹
&#20859;&#27542;&#33337;	Culture vessel	养殖船
&#20859;&#27542;&#33391;&#31181;	Breeding improved varieties	养殖良种
&#20859;&#27542;&#22823;&#33777;&#40070;	Cultured turbot	养殖大鲆
&#20859;&#27542;&#20892;&#25143;	Aquaculture farmers	养殖农户
&#20859;&#27542;&#30149;&#21407;&#20307;	Breeding of pathogens	养殖病原体
&#20859;&#27542;&#24037;&#20316;&#24231;&#35848;&#20250;	Work culture forum	养殖工作座谈会
&#20859;&#27542;&#24687;	Farming income	养殖息
&#20859;&#27542;&#39640;&#20135;&#39640;&#25928;	Breeding high yield and high efficiency	养殖高产高效
&#20859;&#27542;&#32463;&#27982;&#25928;&#30410;	Economic benefits of aquaculture	养殖经济效益
&#20859;&#27542;&#32599;&#38750;&#40060;	Tilapia culture	养殖罗非鱼
&#20859;&#27542;&#34691;&#34809;	Breeding crabs	养殖螃蟹
&#22823;&#27700;&#20135;&#20859;&#27542;&#25143;	Large aquaculture households	大水产养殖户
&#27700;&#20135;&#21697;&#28040;&#36153;	Consumption of aquatic products	水产品消费

### 3.5. Détection des signaux faibles

Pour détecter les signaux faibles, nous recherchons tout d’abord les mots-clés et les termes techniques connus et nous les codons dans le texte (Titre et Résumé). Puis, nous détectons les nouvelles séquences dépassant un seuil d’occurrence fixé par l’utilisateur. Nous croisons ensuite ces nouvelles séquences (les multi-termes) avec le temps et nous ne gardons que celles qui émergent fortement dans la dernière période (ici 2007). Nous croisons enfin ces termes entre eux et nous trions la matrice obtenue par blocs diagonaux. Chaque bloc représente un concept émergent identifié par une terminologie nouvelle non présente dans les thésaurus du domaine et issue seulement de quelques documents récents. Ce sont les fameux « signaux faibles » que nous devons ensuite valider en les croisant avec tous les autres champs de la base et notamment les mots-clés. Dans la figure 11, nous remarquons, le long de la diagonale de la matrice, un certain nombre de clusters constitués de nouveaux termes n’appartenant pas aux mots-clés et qui présentent une organisation sémantique très marquée. Chacun d’eux est alors extrait dans une sous-matrice carrée et visualisé sous forme de graphe sémantique (figure 12). Cette information est alors soumise à l’expert du domaine pour validation.

&#37197;&#21512;&#39282;&#26009;&#20859;&#27542;	Aquaculture feed	配合饲料养殖
&#20859;&#27542;&#22616;	Breeding pond	养殖塘
&#20859;&#27542;&#21697;	Aquaculture products	养殖品
&#20859;&#27542;&#33258;&#36523;&#27745;&#26579;	Breeding self-pollution	养殖自身污染
&#20859;&#27542;&#21487;&#25345;&#32493;&#21457;&#236	Sustainable development of aquaculture	养殖可持续发展
&#20859;&#27542;&#40077;	Abalone aquaculture	养殖鲍
&#20859;&#27542;&#24322;	Culture differences	养殖异
&#20859;&#27542;&#24773;	Culture conditions	养殖情
&#20859;&#27542;&#25345;&#32493;&#20581;&#24247;&#214	Sustained and healthy development of aquacu	养殖持续健康发展
&#20859;&#27542;&#36136;&#37327;&#23433;&#20840;	The quality and safety culture	养殖质量安全
&#20859;&#27542;&#36896;	Culture building	养殖造
&#20859;&#27542;&#27827;&#34809;	Breeding crab	养殖河蟹
&#20859;&#27542;&#33829;&#20859;	Aquaculture Nutrition	养殖营养

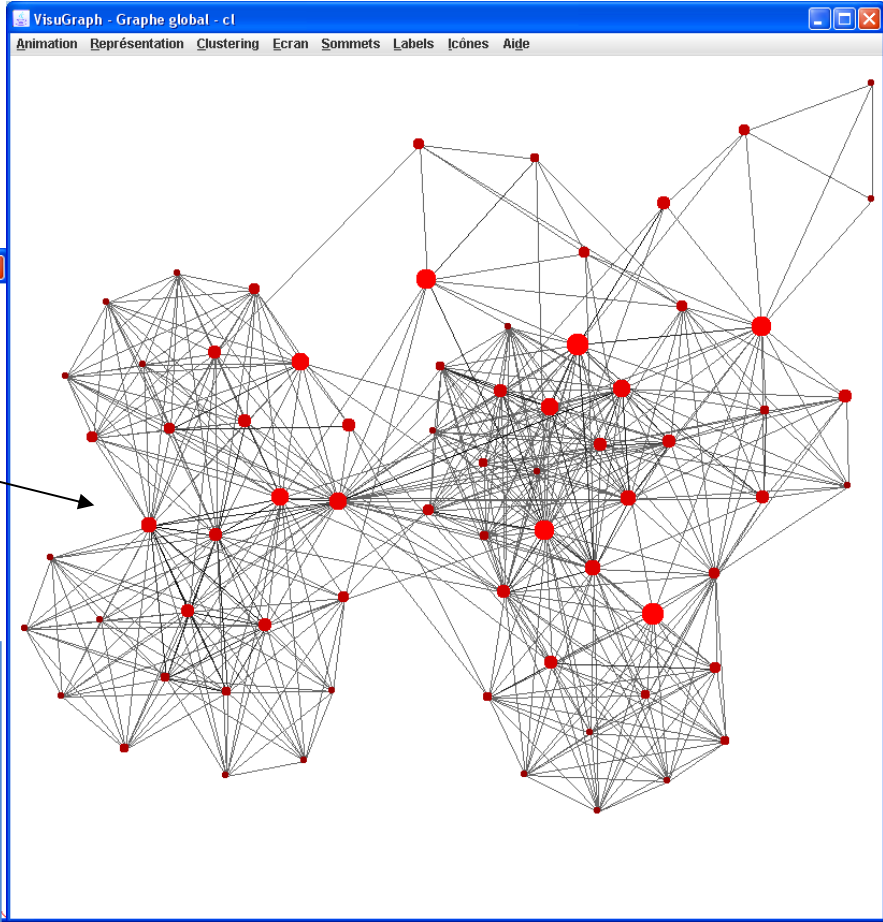
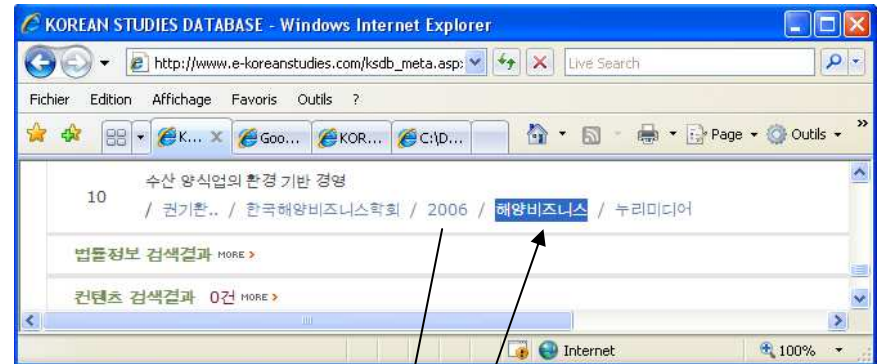


Figure 11 : traduction Google de segments répétés et clusters émergents

Figure 12 : exemple d’un réseau sémantique de multi-termes

## 4. Autres possibilités d'analyse

Voici deux autres exemples de sources dont l'analyse peut être menée par une méthode analogue à celle suivie pour le chinois. L'UNICODE UTF-8 peut être extrait depuis le code source des pages html. Pour Al Jazeera, l'originalité est de pouvoir analyser les réactions des internautes aux articles via le blog. Pour la base en coréen, nous voyons que la plage des caractères réservée à cette langue est différente, mais le principe d'analyse reste le même. A chaque fois, la difficulté est de trouver un balisage suffisant permettant de catégoriser au mieux les informations avant analyse (acteurs, sémantique, dates, ...). Des dictionnaires de mots-clés et d'expressions sont aussi très utiles pour traiter le texte libre et y détecter l'innovation.



Idéogramme d'un terme coréen et code UTF-8 correspondant



Figure 13 : le site d'Aljazeera.net (article bref et blog correspondant)

Figure 14 : cas du site www.e-koreanstudies.com

## 5. Conclusion

La base documentaire CQVIP, sur laquelle nous avons réalisé nos tests, ne représente qu'un exemple des multiples sources qu'il est possible de traiter par cette méthode, aussi bien pour le chinois que pour d'autres langues comme le japonais, le coréen, l'arabe. Il reste cependant de nombreux problèmes à résoudre, pour que cette démarche devienne parfaitement opérationnelle.

- Etablir des dictionnaires thématiques et leur traduction en anglais (et/ou en français).
- Traiter les entités nommées (pour les auteurs ou certains organismes ou journaux la traduction automatique est suffisante), mais il reste de nombreuses ambiguïtés à lever (influence des accents, de la prononciation, du contexte).
- La traduction de la nouvelle terminologie détectée de façon statistique ne sera pas nécessairement présente ni dans les dictionnaires thématiques car trop récente ou tout simplement inconnue des non initiés, ni dans la traduction proposée par *Google* pour les mêmes raisons.

Il devient donc nécessaire à terme d'envisager une collaboration entre experts :

- de la fouille de texte
- du traitement automatique des langues (sémantique, morphosyntaxique, ontologies, ...)
- des langues (chinois, coréen, japonais, arabe, ...)
- des domaines à analyser (scientifique, technologique, économique, géopolitique, ...)

Cette collaboration peut intervenir à deux niveaux : la préparation des données (homogénéisation du vocabulaire, choix des granularités, traductions, désambiguïsation, ...) et l'interprétation des résultats, car bien souvent nous sommes obligés de revenir aux documents sources qui comportent des passages assez importants en texte libre, qu'il vaut mieux appréhender dans la langue d'origine.

## 6. Bibliographie

- [1] **DOUSSET B.**, *Extraction of strategic information through analysis of major components* . Datametrics Journal, volume 2, april 2008, issue 1.
- [2] **GHALAMALLAH I., GRIMEH A., DOUSSET B.**, *Processing data stream by relational analysis*. MODULAD, INRIA, n°36, july 2007.
- [3] **GHALAMALLAH I., LOUBIER E., DOUSSET B.**, *Business intelligence a proposal for a tool dedicated to the analysis relational*. International Journal of Competitive Intelligence, Strategic, Scientific and Technology Watch, SciWatch Journal, hexalog, Barcelona - Spain, Vol. 3, august 2008.
- [4] **GUENEC N., LOUBIER E., GHALAMALLAH I., DOUSSET B.**, *Management and analysis of chinese database extracted knowledge*. Flexible Query Answering (FQAS 2008), Londres , september 22, 2008.
- [5] **LOUBIER E., DOUSSET B.**, *Visualisation and analysis of relational data by considering temporal dimension*. International Conference on Enterprise Information Systems (ICEIS 2007), Funchal, Madeira - Portugal, Vol. ISAS, INSTICC Press, p. 550-553, june 12-16 2007.
- [6] **ROUX C.**, *Methodes of sort for extracting weak signals*. International Journal of Competitive Intelligence, Strategic, Scientific and Technology Watch, SciWatch Journal, hexalog, Barcelona - Spain, Vol. 2, april 2009, issue 1, pp 23-29.